DOCUMENT RESUME

ED 392 832 TM 024 524

AUTHOR Kennedy, Rob
TITLE A Comparison of Limited vs. Unlimited Retakes of a Multiple-Choice Test.
PUB DATE 10 Nov 94
NOTE 10p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Nashville, TN, November 9-11, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Graduate Students; Graduate Study; Higher Education; Item Banks; *Multiple Choice Tests; Research Design; *Scores; *Student Motivation; Test Items; *Test Results
IDENTIFIERS *Test Repeaters

ABSTRACT

The purpose of this study was to compare the scores of students who were allowed unlimited retakes of a multiple-choice test with the scores of students who were limited to only four retakes (five trials) of the same test. The tests were each made up of 20 randomly drawn questions from a large pool of questions about research methods. Three graduate research classes were involved in the study; two were the limited groups (11 and 18 students) and one was the unlimited group (14 students). The groups were analyzed using a 3 x 5 repeated measures analysis of variance design. The between-subjects factor was treatment group and the within-subjects factor was test administration (five administrations). The results indicated significant differences ($p<0.01$) both between and within groups. The interaction effect was not significant. Additional research is necessary to clarify the appropriate number of trials for tests to insure adequate learning while providing sufficient motivation for students to study rather than memorize answers. (Contains 13 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ROBERT L. KENNEDY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 392 832

A Comparison of Limited vs. Unlimited Retakes of a Multiple-Choice Test

Rob Kennedy
Department of Educational Leadership
University of Arkansas, Little Rock

Mid-South Educational Research Association Annual Meeting
Maxwell House
Nashville, Tennessee
November 10, 1994

TM024524

BEST COPY AVAILABLE

Abstract

The purpose of this study was to compare the scores of students who were allowed unlimited retakes of a multiple-choice test with the scores of students who were limited to only four retakes (five trials) of the same test. The tests were each made up of twenty randomly drawn questions from a large pool of questions about research methods. Three graduate research classes were involved in the study; two were the limited groups and one was the unlimited group. The group sizes for which there was complete data were 11, 18, and 14, respectively. The groups were analyzed using a 3 x 5 repeated measures analysis of variance design. The between-subjects factor was treatment group (limited groups and unlimited group) and the within-subjects factor was test administration (five administrations). The results indicated significant differences ($p<0.01$) both between and within groups. The interaction effect was not significant.

A Comparison of Limited vs. Unlimited Retakes

of a Multiple-Choice Test

If practice makes perfect, then the opportunity to retake tests should lead to improved performance in the classroom. In fact, tests are retaken routinely in mastery learning environments (Caponigri & Schumann, 1982) and many other situations. For example, Knight (1973) incorporated retakes in a "programmed achievement" approach to insure that students reached mastery before they were allowed to continue to subsequent lessons. Van Winkle (1978) actually required retakes rather than making them an option. In an interdisciplinary (chemistry, physics, and biology) science course for nonscience majors at the University of Michigan-Dearborn, students were given nine quizzes consisting of ten true-false and ten matching questions. If students did not reach a ninety percent passing level, then they were required to retake the quiz at the beginning of the next lab session. Average performance on the tests was greatly increased due to the retakes. In fact, only one percent had a final quiz total lower than their pre-retake total.

John, Ruminski, and Hanks (1991) surveyed journalism educators for their admission requirements and determined that the majority of the respondents allowed retakes of the entrance exam, with responses ranging from no retakes (4 of 86); to one (12), two (17), or three (12) retakes; to unlimited (28). Thirteen of the respondents either specified other conditions or did not provide a response to the question of whether retakes were possible.

Lore-Lawson (1993) reported using exam retakes at Cardinal High School in Eldon, Iowa, as a self-esteem builder. She splits the difference with students who retake tests: "Why punish kids for learning? I learned more from items I got wrong on tests than I ever did from most textbooks. Many of my students appreciate this policy and me for having it (p. 2-3)".

Not all retake reports were as positive as those already mentioned. For instance, Stoker and Parker (1976) allowed students in an introductory-level college chemistry class the opportunity to retake tests to improve their scores. Although one in four students did improve their scores, the majority of students who were willing to make the effort were the students who were making C's, D's, and F's. It is probably not surprising that only a fourth of them improved due to the retakes.

Araujo and Semb (1979) investigated order of items on student performance under the Keller Plan and under a Contingency Managed Lecture approach devised by Semb and a colleague. One offshoot of the study was a consideration of the effect of retakes on performance. In neither case was a significant improvement found from allowing a retakes of exams. In the case of the Personalized System of Instruction (Keller) approach the authors concluded that the method was strong enough that retakes were not particularly needed.

In some cases, allowing retakes led to drawbacks. Elbrink (1973), for example, observed improvement from retakes among freshmen enrolled in one of two calculus sequences in the CRIMEL (Curriculum Revision and Instruction in Mathematics at the Elementary Level) project at The Ohio State University. Although statistical evidence was not provided in the report of the study, Elbrink stated that the median and mean scores increased significantly between attempts. However, he further observed that students did not take each attempt seriously because of the opportunity to take unlimited retakes. Some of the students viewed only their last retake as "the" test. In response to this situation, Elbrink planned to allow only one retake in later studies.

Davik (1980) noted a similar problem with students in a high school chemistry class. By allowing unlimited retakes, students appeared to be "willing to take their chances on a test, without proper review and study (p.213)". Davik's response was to allow a maximum of a raise

to a C grade after a retake. He reported that requests for retakes had dropped substantially after the change in procedure.

This researcher has observed a similar phenomenon with graduate students retaking tests in an introduction to research methods class. Some of the students study before tests and need only a few retakes while others have been documented as taking as many as eighteen tests before reaching 100 percent mastery (although only 90 percent was required). In those cases, it seemed clear that those students were simply memorizing answers to randomly-selected multiple-choice questions rather than studying, which of course, defeated the purpose of the retakes. Although Elbrink and Davik proposed allowing only one retake for a total of two tests, it was proposed for this study that four retakes (five trials) be allowed. Little literature was found to support this figure, although Karp (1983) reported that University of Houston at Clear Lake City users of the Keller Plan indicated using retakes of one, two, three, five, and unlimited. Based on empiracle observation, however, most students who appeared to be prepared for the research methods tests required only a few retakes, so five was selected as a fair number of opportunities for serious students to be able to succeed. In a previous mastery learning environment coordinated by this researcher, five tests were available for students, but were rarely exhausted, further supporting the selection of five trials.

The purpose of this study, then, was to compare the first five scores of students who were allowed unlimited retakes of a multiple-choice test with the scores of students who were limited to only five trials (four retakes) of the same test. The tests were made up of twenty randomly drawn questions from a large pool of questions about research methods. Three graduate research classes were involved in the study; two were the limited groups and one was the unlimited group. The group sizes for which there was complete data were 11, 18, and 14, respectively. The

groups were analyzed using a 3 x 5 repeated measures analysis of variance design. The between-subjects factor was treatment group (the two limited groups and the one unlimited group) and the within-subjects factor was test administration (five administrations).

The assumptions for the two-factor, repeated-measures (mixed design) study include those for independent groups and single-factor repeated measures designs (Girden, 1992). That is, within-group variability should be equal across groups, and the scores should be normally distributed and independent among groups. In addition, the population variance-covariance matrices should be equal and their pooled matrix should have a sphericity pattern.

The Statistical Package for the Social Sciences (SPSS) computer program was used for data analysis. The three groups were tested for within-group variability. There was insufficient evidence to reject the null hypotheses that there were no differences among the variances, using Levene statistics. It was assumed, then, that the homogeneity assumption was met.

The test scores were tested for normality. The Lilliefors test (a modification of the Kolmogorov-Smirnov test) indicated outright normality or sufficient normality among the groups to satisfy the normality assumption. The robust quality of the analysis of variance allows acceptably accurate interpretations with the small departures observed. The tenability of independence seems reasonable since the subjects tested did so individually and did not affect the scores of the others.

The variables' pooled matrix should display a sphericity pattern (Girden, 1992). SPSS provides the Mauchly test of sphericity for this purpose. The statistic for the test scores was 0.61, with a significance level of 0.018, indicating that the null hypothesis of no relationship could be rejected. It was concluded that the dependent variables were related.

The variance-covariance matrices were computed for the test scores. Two homogeneity-of-variance tests, Cochran's C and the Bartlett-Box F, are computed in SPSS, and were applied to each variable. The tests all yielded probabilities at 0.05 or above, suggesting that there was insufficient evidence to reject the null hypotheses that the variances were equal. Box's M, based on both the determinant of the variance-covariance matrices and the pooled variance-covariance matrix, provided a multivariate test for the homogeneity of the matrices. Since Box's M test is very sensitive to departures from normality, the significance level can be based on both F and chi-square statistics. Since some departure from normality was indicated in the data, the chi-square-based statistic is reported here: 20.74 with an approximate (as reported by SPSS) probability of 0.90. Given this level, there is insufficient evidence to reject the null hypothesis that there is no difference in the variances of the variance-covariance matrices.

The repeated measures analysis of variance table indicated significant differences both within and between groups. The between group significance was 0.0063 and the within group significance was 0.0000 ($p<0.00005$). The means for the two limited-retake groups were 13.1 and 12.7 while the mean of the unlimited-retake group was 14.3. Within groups the means ranged from 10 to 14.9 for the limited groups, but from 11.9 to 16.2 for the unlimited groups. Although there are noticeable differences between the two group types, the gap is only a point or two. Strangely, the group with unlimited trials had the highest average. It is possible that the pressure of having a limited number of trials affected the outcome. It is also possible that the contamination caused by random selection of items could have biased the scores. In fact, that problem has been discussed by Sarvela and Noonan (1987). The small sample sizes involved in this study could also be a limitation of the study. Further studies with larger samples would certainly be appropriate.

Given that unlimited amounts of time for tests is not always practical, particularly in situations where some students must leave for classes or other reasons and cannot benefit from the extended time, it is worthwhile to arrive at some indication of an appropriate number of trials for tests to insure adequate learning while providing sufficient motivation for students to study rather than memorize answers. While five trials has been suggested here, it is suggested that more than one opportunity be provided if possible. This study and the literature support the benefits of repeated testing.

References

Araujo, J., & Semb, G. (1979, September). The Effects of Item Order on Student Performance. Paper presented at the Annual Meeting of the American Psychological Association, New York, NY. (ERIC Document Reproduction Service No. ED 185 086)

Caponigri, R. S., & Schumann, M. R. (1982). Mastery Learning Workbook 1982. (ERIC Document Reproduction Service No. ED 217 941)

Davik, J. (1980, March). View from My Classroom: Discovering Chemistry and Some Departmental Comments. Journal of Chemical Education, 57(3), 213-15.

Elbrink, L. C. (1973). A Flexible and Forgiving Testing Program: A Report of the Current Status of the Testing Program of the CRIMEL Project. School Science and Mathematics, 73(8), 686-90.

Girden, E. R. (1992). ANOVA: Repeated Measures. (Sage University Paper series on Qualitative Applications in the Social Sciences, series no. 07-084). Newbury Park, CA: Sage.

John, J. A., Ruminski, H., & Hanks, W. (1991, Fall). Trends in Writing Skills Tests for Admission to Programs. Journalism Educator, 46(3), 44-49.

Karp, H. J. (1983, April). The Use of Keller's Personalized System of Instruction. Paper presented at the Annual Meeting of the Southwestern Psychological Association, Dallas, TX. (ERIC Document Reproduction Service No. ED 231 295)

Knight, J. M. (1973). The Effect of Programmed Achievement on Student Performance. Journal of Educational Research, 66(7), 291-94.

Lore-Lawson, J. (1993). Self-Esteem in the Foreign Language Classroom: Activities That Do More Than Teach. (ERIC Document Reproduction Service No. ED 369 267)

Sarvela, P. D., & Noonan, J. V. (1987, April). Testing and Computer-Based Instruction: Psychometric Considerations. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, D.C. (ERIC Document Reproduction Service No. ED 282 897)

Statistical Package for the Social Sciences. SPSS, Inc. 444 N. Michigan Avenue. Chicago, IL 60611.

Stoker, H. S., & Parker, O. R. (1976). Allowing Students a Second Chance on Examinations. Journal of College Science Teaching, 5(4), 232-35.

Van Winkle, L. J. (1978, November). Remedial Study Plus Retake Exams Equals Better Grades. Journal of College Science Teaching, 8(2), 91-92.